Analyzing the Empirical Effectiveness of COVID-19 Vaccines in Reducing the Transmission Rate on Different Variants in the United States

Author: George Zhang

Summary of Research Questions and Results

1. What is the distribution of COVID-19 vaccine administrations across the United States?

- The West coast and the Northeast states have the highest percentage of fully vaccinated individuals along with some inner states such as Colorado and New Mexico.
- The percentage of fully vaccinated individuals in each state ranges from around 50% to around 80%.

2. What is the distribution of each type of COVID-19 vaccine (Pfizer/Moderna/J&J) across the United States? How do they correlate with the distribution of all COVID-19 vaccine administrations? Which type of vaccine is the most popular?

- The Pfizer vaccine is the most popular, then Moderna, and J&J is the least popular.
- The plot for the distribution of fully vaccinated individuals by Pfizer and Modern look fairly like the distribution of all vaccinated individuals.
- Most states have a low percentage of individuals fully vaccinated by J&J. However, Maine has the highest percentage at around 10%, more than any other state.

3. How has the transmission rates changed in response to the increased vaccination rates in the United States? For Washington State specifically?

- The introduction of vaccines aligned with a significant reduction in the transmission rate.
- The appearance of the Delta variant increased the transmission rate, which went down as the percentage of fully vaccinated individuals continued to increase.
- Something similar occurred with the introduction of the Omicron variant.

4. How accurately can the number of fully vaccinated individuals be used for predicting new COVID-19 cases?

- I initially used population, state area, and total COVID cases as features to predict new COVID-19 cases. Then I added the number of fully vaccinated individuals as a feature.
- Neither case was able to produce a good prediction. Including fully vaccinated individuals as a feature also did not increase the accuracy of the model.

Motivation

COVID-19 is arguably one of the greatest challenges that the world has ever had to face, and vaccines are one of the tools to combat COVID-19. But are COVID-19 vaccines effective in reducing transmission rates? We often hear about the effectiveness of vaccines through tests and trials in the media, but is that reflected in the real world? Being able to provide empirical and real-world information on the effectiveness of COVID-19 vaccines can help reduce mistrust and misinformation surrounding the vaccine. Additionally, COVID-19 has undergone multiple mutations, namely the delta and the omicron variants since it began spreading. Obtaining a deeper understanding of the effects of COVID-19 vaccines on each variant can aid scientists in future vaccine research.

Dataset

COVID-19 Vaccinations in the United States,Jurisdiction

- Export > CSV

United States COVID-19 Cases and Deaths by State over Time

- Export > CSV

United States Cartographic Boundary

- Open the side bar > right click “gz_2010_us_040_00_5m.json” > Download

United States Population by State

- Downloads automatically

Python Dictionary of US States and Territories

- No download necessary. I have already added the dictionary to my code.

Method

Setup

1. Import “United States Cartographic Boundary” dataset (Challenge Goal: **Multiple Datasets**).
2. Import “United States Population by State” dataset (Challenge Goal: **Multiple Datasets**).
3. Merge state boundaries dataset with the population dataset to create state data.
4. Create dictionary mapping state abbreviation to full names using the “Python Dictionary of US States and Territories” dictionary.
5. Import “COVID-19 Vaccinations in the United States,Jurisdiction” dataset.
6. Add full state name column to the vaccination dataset.
7. Merge state data with the vaccination dataset.
8. Repeat steps 5-7 for the “United States COVID-19 Cases and Deaths by State over Time” dataset.

1. What is the distribution of COVID-19 vaccine administrations across the United States?

1. Calculate the percentage of fully vaccinated individuals by dividing the count by the population.
2. Plot percentage of fully vaccinated individuals of each state using geospatial data.

3. The question can be answered by looking at the map of the total vaccinations of each state.

2. What is the distribution of each type of COVID-19 vaccine (Pfizer/Moderna/J&J) across the United States? How do they correlate with the distribution of all COVID-19 vaccine administrations? Which type of vaccine is the most popular?

1. Calculate the percentage of fully vaccinated individuals for each type of vaccine by dividing the count by the population.
2. Make multiple subplots of the percentage of fully vaccinated individuals of each type of vaccine using geospatial data.
3. We can see the distribution of each type of vaccine by looking at the map of each type of vaccine for each state.
4. We can see the correlation with the distribution of all COVID-19 vaccine administrations by comparing this map with the previous map of total vaccinations.
5. For each state, find which type of vaccine has the most fully vaccinated individuals. Plot the vaccine that is most popular in each state.
6. We can which vaccine is the most popular by looking at the plot above.

3. How has the transmission rates changed in response to the increased vaccination rates in the United States? For Washington State specifically?

1. Extract a week's worth of data at the following times
    a. Alpha variant (before vaccines)
    b. Alpha variant (after vaccines)
    c. Delta variant (earlier)
    d. Delta variant (later)
    e. Omicron variant (earlier)
    f. Omicron variant (later)
2. For each time segment, plot the
    a. Cases per population
    b. Deaths per population
    c. Mortality rate (deaths/cases)
    d. Transmission rate (new cases/total cases)
    e. Fully vaccinated individuals per population
3. We can see how the transmission rate changes as vaccination increases and for different variants.
4. For Washington State, for each individual time range (see step 1) in the entire time range
    a. Calculate the $\log_{10}$ of new cases and total cases
    b. Add the scatter plot of the new cases versus total cases
    c. Include a legend and a different color if applicable for the individual time range
5. We can see how the transmission rate has deviated from exponential growth (straight line with a positive slope) at different times and different variants.

4. How accurately can the number of fully vaccinated individuals be used for predicting new COVID-19 cases?

1. import scikit-learn, specifically the DecisionTreeRegressor and MLPRegressor models (Challenge Goal: **Machine Learning**).
2. Separate features (number of total cases, population, census area) and label (number of new cases). Have another set of data to include the number of fully vaccinated individuals in the features.
3. Create DecisionTreeRegressor and MLPRegressor
4. For 100 iterations
    a. Separate into training (80%) and test (20%) sets.
    b. Train model on training data without vaccination data.
    c. Make predictions on test data.
    d. Calculate accuracy using mean_squared_error (for Decision Tree) and MLPRegressor.score (for MLP).
    e. Print out the accuracy of both models.
    f. Repeat steps b-e using training data with vaccination data.
5. See how accurate the models are in general and whether including the vaccination data improved the accuracy a majority of times out of 100 times.

Results

1. What is the distribution of COVID-19 vaccine administrations across the United States?

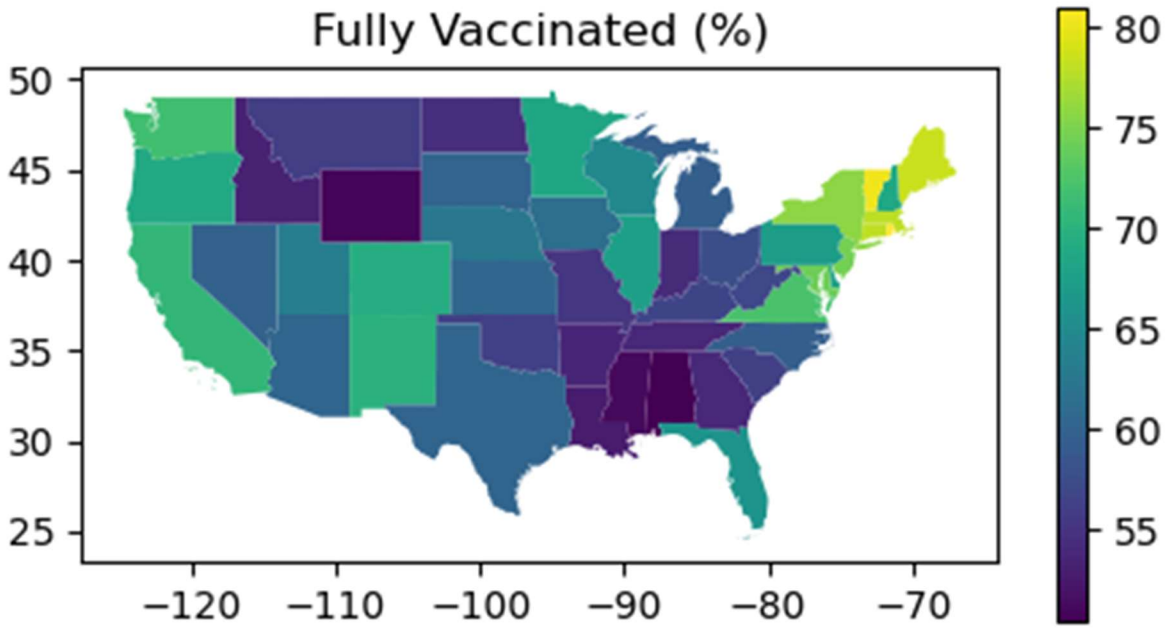
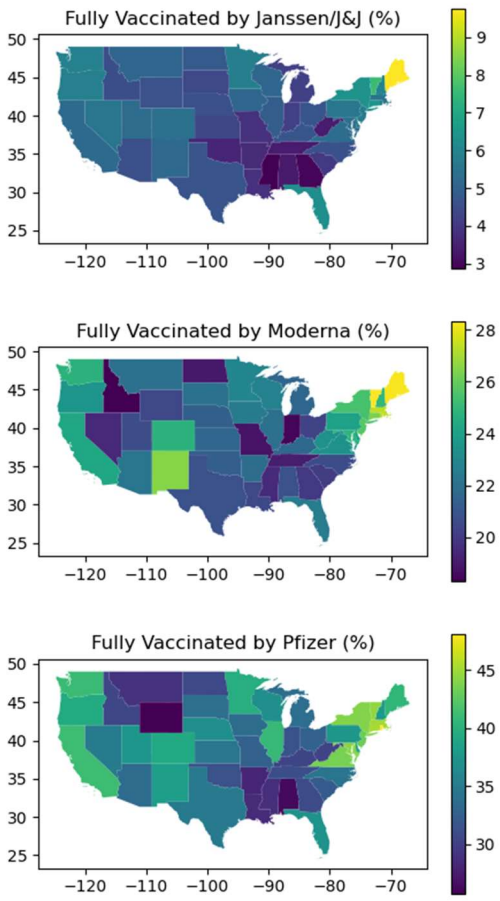


Fig 1. The Distribution of Individuals Fully Vaccinated for COVID-19 as of 3/11/2022.

Figure 1 shows the distribution of individuals who are fully vaccinated for COVID-19 (two doses of Pfizer or Moderna, or one dose of J&J). The West coast and the Northeast states have the highest percentage of fully vaccinated individuals along with some inner states such as Colorado and New Mexico. This distribution looks like the distribution of democratic versus republican states in the U.S (Red states and blue states - Wikipedia). Therefore, the distribution of COVID-19 vaccines is at least in part correlated with political parties. However, that is not to disregard the individual choices to become vaccinated in even states with low vaccination rates. Additionally, the majority of the people in the United States are fully vaccinated regardless of the state they belong to. The percentage of fully vaccinated individuals in each state ranges from around 50% to around 80%.

2. What is the distribution of each type of COVID-19 vaccine (Pfizer/Moderna/J&J) across the United States? How do they correlate with the distribution of all COVID-19 vaccine administrations? Which type of vaccine is the most popular?



The Pfizer vaccine appears to be the most popular with around 25% to 50% of individuals in each state fully vaccinated by the Pfizer vaccine. Moderna is the second most popular with around 15% to 30% of individuals in each state fully vaccinated by the Moderna vaccine. J&J is the least popular with around 3% to 10% of individuals in each state fully vaccinated by the J&J vaccine.

The plot for the distribution of fully vaccinated individuals by Pfizer and Modern look fairly like the distribution of all vaccinated individuals. This means that no state has an extreme preference for either the Moderna or the Pfizer vaccine.

Most states have a low percentage of individuals fully vaccinated by J&J. However, Maine has the highest percentage at around 10%, more than any other state. Interestingly, Maine also seems to prefer the Moderna vaccine, along with the J&J vaccine, over the Pfizer vaccine compared to other states. For example, Maine and Washington have a similar percentage of individuals fully vaccinated by Pfizer but Maine has a higher percentage of individuals fully vaccinated by Moderna and J&J.

Fig 2. Distribution of different vaccine manufacturers.

Looking at the most popular vaccine manufacturer across the United States (Fig 3), we can see that all states prefer Pfizer over any other vaccine. In conclusion, there does not appear to be any significant preferences for vaccines produced by different manufacturers other than the overall preference for the Pfizer COVID vaccine.

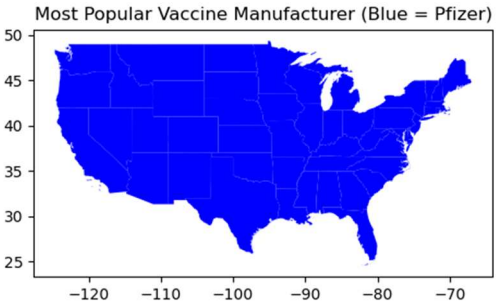


Fig 3. Most popular vaccine manufacturer.

3. How has the transmission rates changed in response to the increased vaccination rates in the United States? For Washington State specifically?

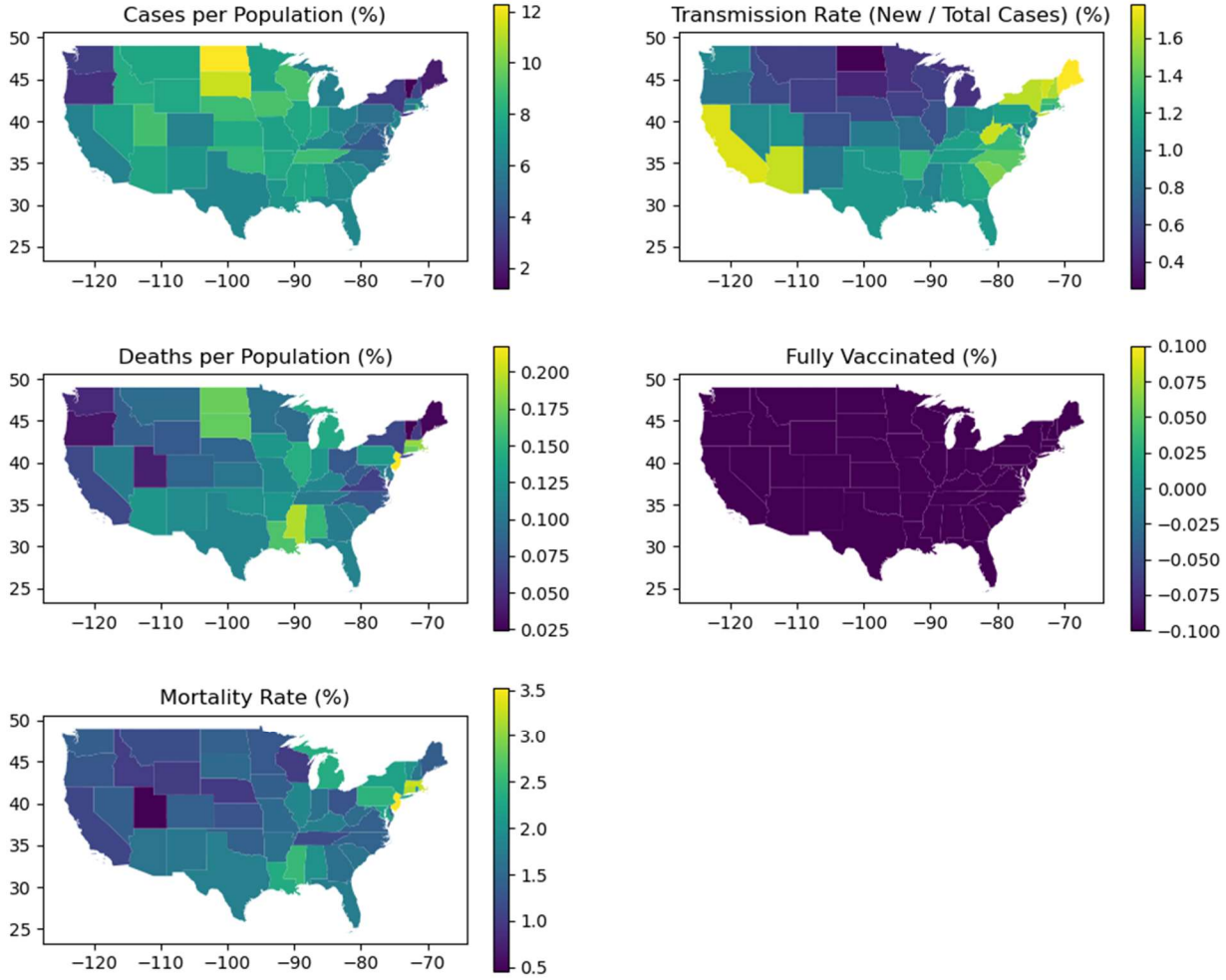


Fig 4. COVID statistics as of January 2021 (alpha variant, no fully vaccinated individuals).

I gathered COVID cases and vaccination data at different times to discover how they have changed and any correlations (I used https://usafacts.org/articles/covid-variants-delta-alpha-common/ to identify which times corresponded to which variant). The first of such times is in January 2021 when vaccines have not been administered and when the alpha variant was the most popular variant. At this point, the transmission rate (number of new cases / numbers of total cases) is at an all-time high at around 0.4% to 1.6% (top right of Figure 4). This means that each day, people equaling to 1% of the number of all cases is catching COVID.

At this time, around 2 to 12% of the population of each state has already gotten COVID. The mortality rate of catching COVID is at around 0.5 to 3.5% for each state.

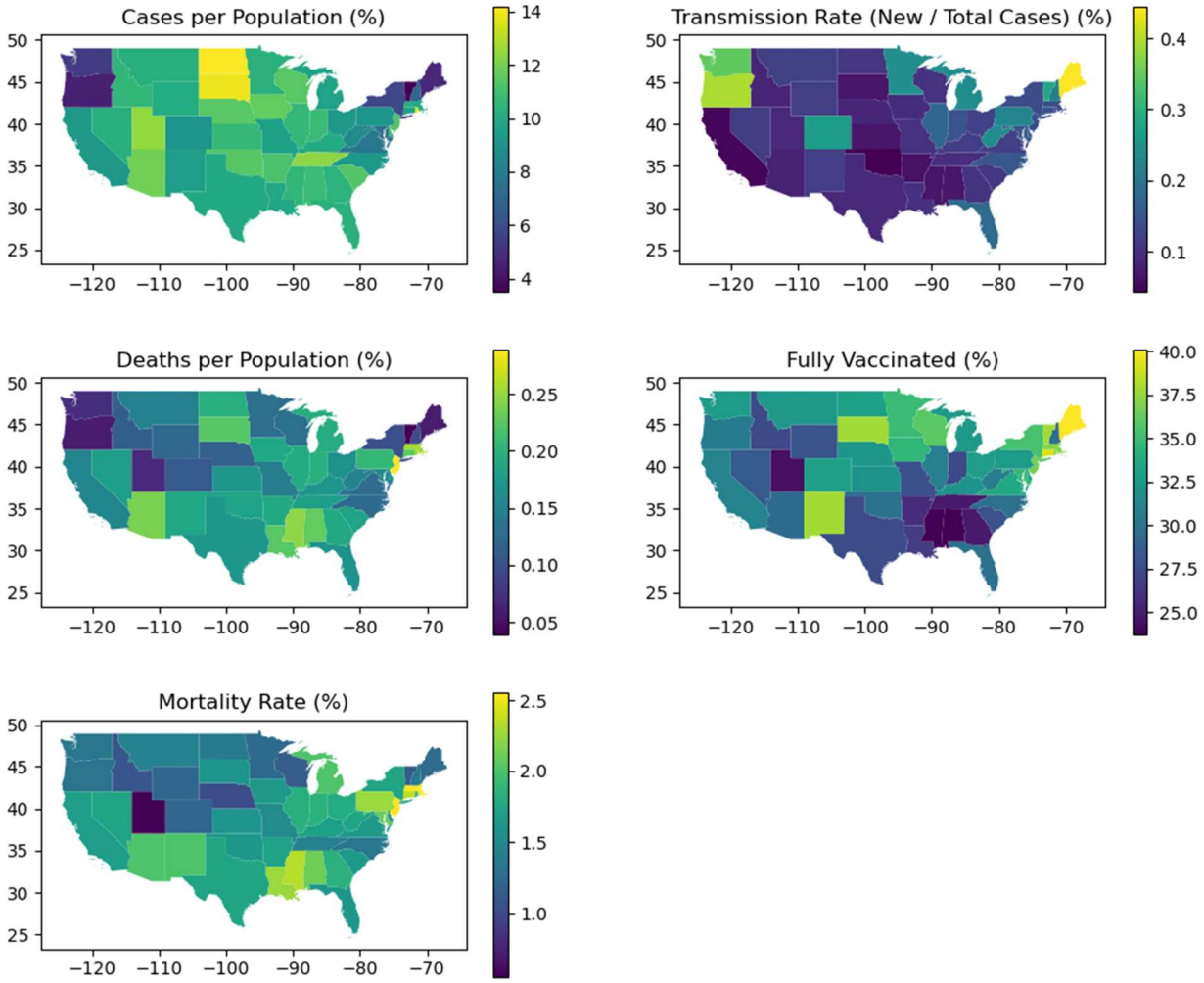


Fig 5. COVID statistics as of May 2021 (alpha variant).

As we reach May 2021, the total number of fully vaccinated individuals have gone up drastically to around 25% to 40% for each state. In correlation, the transmission rate also went down significantly to around 0.1% to 0.4% (compared to 0.4% to 1.6% in January).

Interestingly, the transmission rate is not necessarily correlated to the percentage of fully vaccinated individuals in each state. Maine has the most fully vaccinated individuals at around 40% but also has the highest transmission rate at around 0.4%. However, this is still a significant decrease from 1.6% in January. Since Maine began with a high transmission rate, vaccine could have reduced transmission rate significantly, but still not enough to compare to other states that began with a lower transmission rate.

At this time, COVID cases have increased to 4 to 14% of the population of each state. The mortality rate of catching COVID has lowered slightly to around 0.5 to 2.5% for each state.

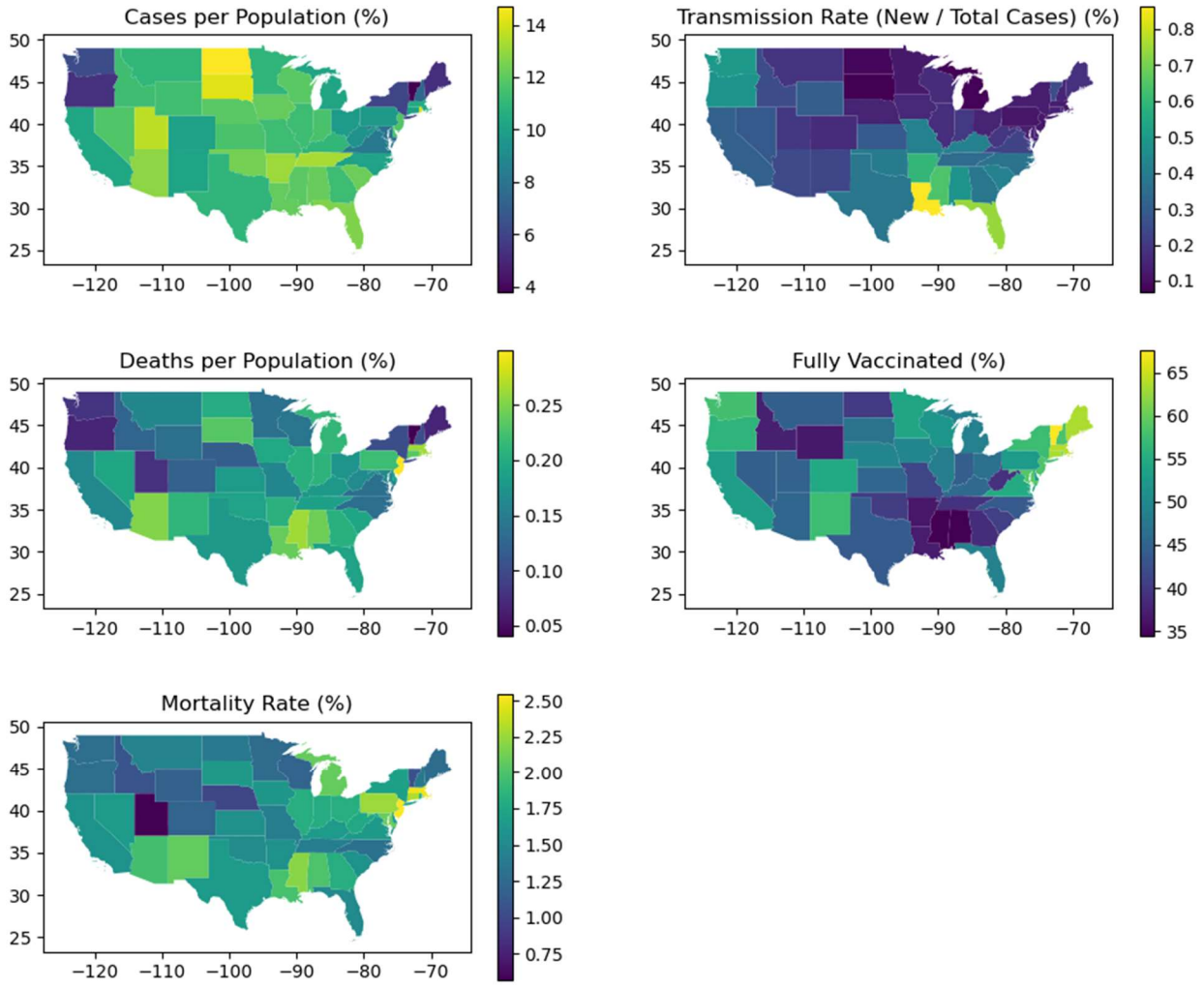


Fig 6. COVID statistics as of August 2021 (Delta variant).

By August 2021, the Delta variant of COVID-19 ha become widespread in the United States. Therefore, despite the percentage of fully vaccinated individuals continuing to rise to around 35% to 65%, the transmission has gone up to around 0.1% to 0.9%. This corresponds to the fact that scientists have discovered the Delta variant to be more transmissible.

By this time, the majority of people in Maine have been fully vaccinated. This can be seen reflected in their extremely low transmission rates. Vaccination is also corresponding more with transmission rate in states such as Louisiana. Louisiana has the highest transmission rate at around 0.9% and a low vaccination at only 40% of fully vaccinated individual.

At this time, COVID cases have increased slightly to 4 to 15% of the population of each state. The mortality rate of catching COVID has remained the same at around 0.5 to 2.5% for each state.

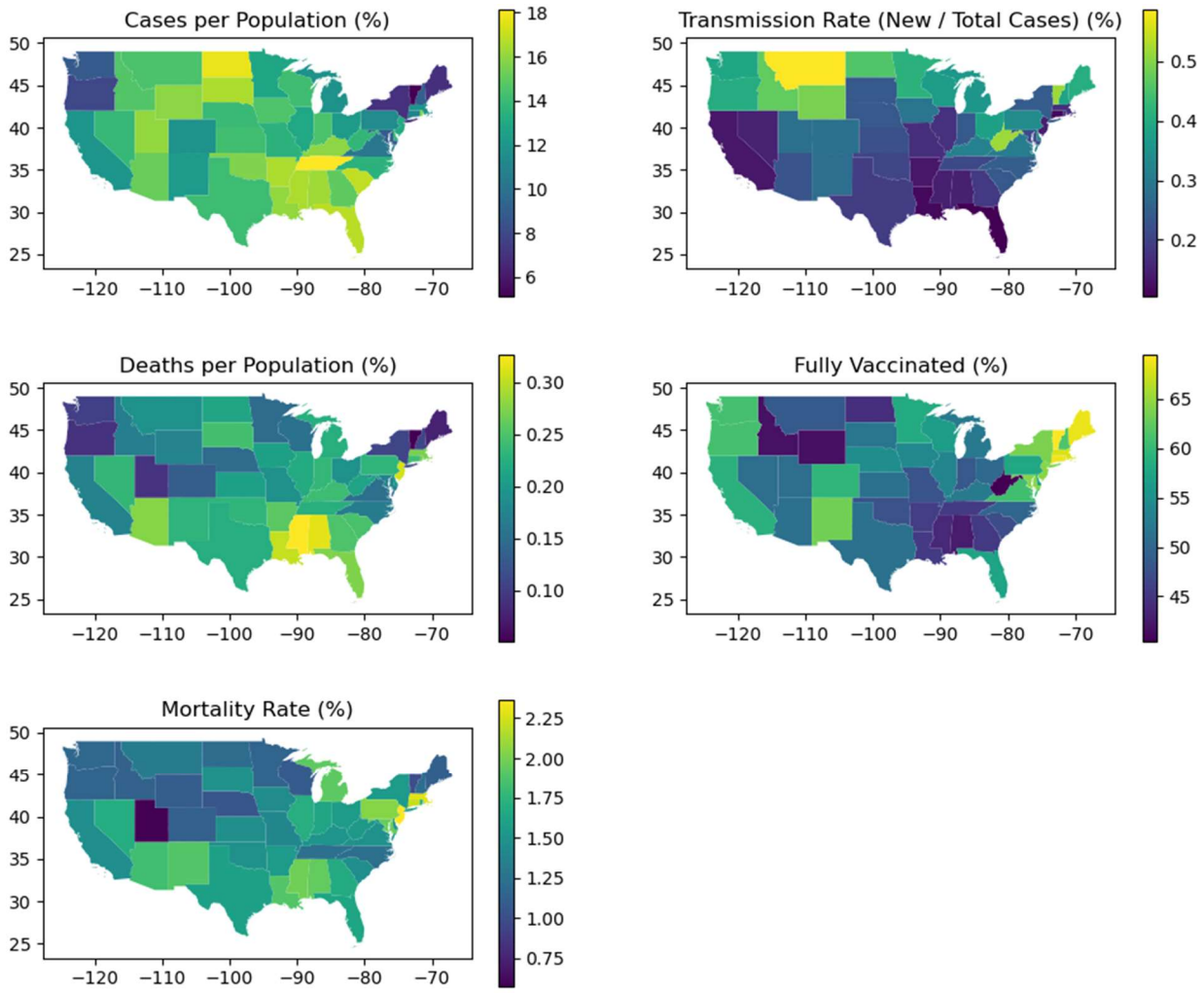
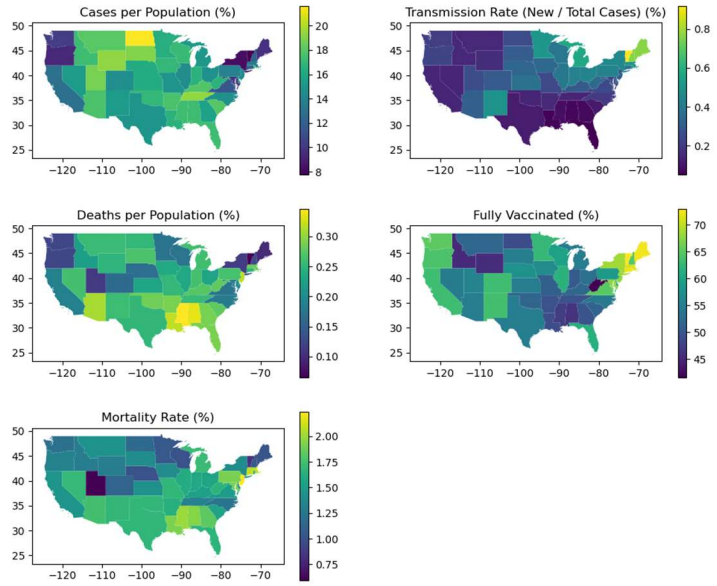


Fig 7. COVID statistics as of October 2021 (Delta variant).

After a few months of the Delta variant, transmission rate has gone down again while states behind in vaccination caught up. The transmission rate decreased to around 0.1% to 0.6% from around 0.1% to 0.8%. Vaccination has gone up from around 35% to 65% to around 40% to 70%. It appears that increased vaccination continued to be effective in reducing transmission rates even for the Delta variant.

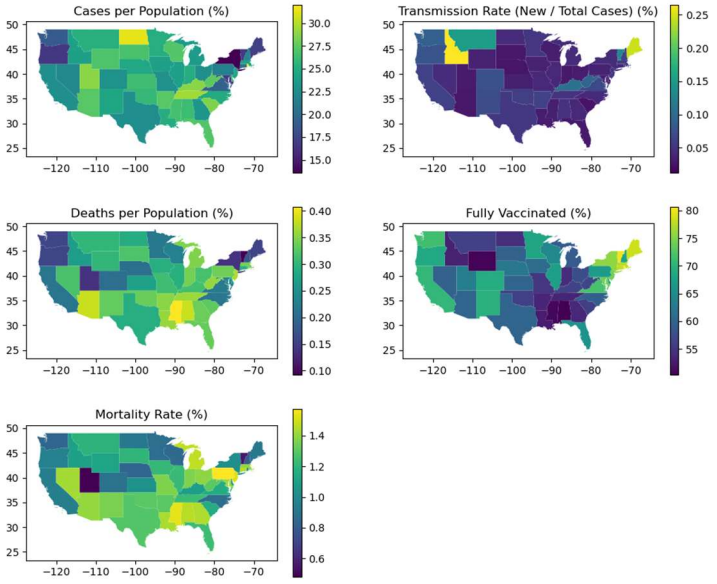
Many states also have corresponding transmission rate and vaccination. For example, Montana has a low vaccination with a high transmission rate. On the other hand, New Mexico has a high vaccination rate and a low transmission rate. However, some states in the South break this pattern because they appear to have low vaccination and low transmission rates as well.

At this time, COVID cases have increased to around 5% to 18% of the population of each state. The mortality rate of catching COVID has remained the same at around 0.5 to 2.5% for each state.



Omicron became the most common variant as of December 2021 and is still the most common variant as of March 2022. A similar trend can be seen with the Omicron variant as with the Delta variant. Due to a higher transmissibility, the transmission rates in the United States initially increased. But as vaccination caught up, the transmission went down again.

Fig 8. COVID statistics as of December 2021 (Omicron variant).



By March of 2022, the COVID cases have increased drastically to around 15% to 30%. The mortality has decreased again to around 0.5% to 1.5%.

Overall, there did appear to be a correlation between higher vaccination and lower transmission rates. This pattern appeared in most states but not all. There can be many explanations to this fact. Because of different individual state regulation, population density, and other factors, vaccination is unable to fully explain the changes in the transmission rates.

Fig 9. COVID statistics as of March 2022 (Omicron variant).

Additionally, the appearances of new variants have increased transmission rates by a significant amount. However, vaccines appear to continue to be effective in reducing transmission rates despite the temporary setback.

The mortality rate did not seem to be affected significantly by different COVID variants and has either stayed the same or decreased. This can be either a result of vaccination, improved knowledge and equipment in caring for COVID patients, or another factor.

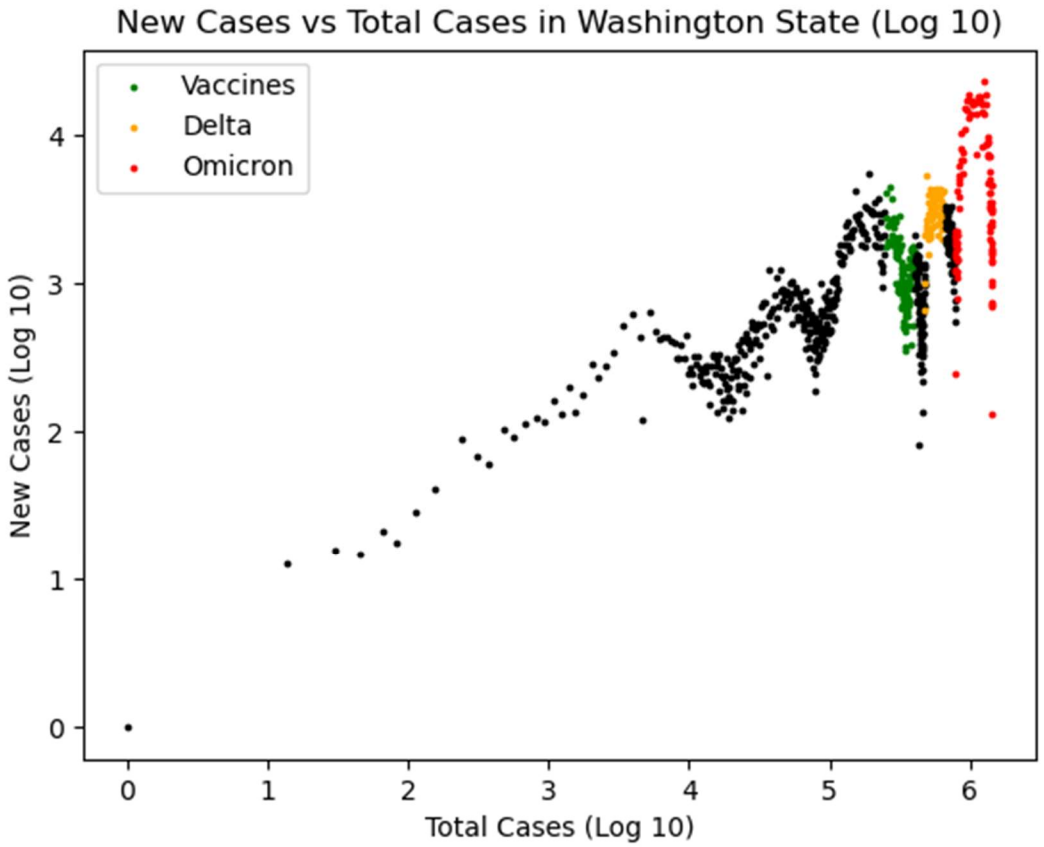


Fig 10. COVID statistics in Washington State from January 2020 to March 2022.

To see the detailed effects of vaccination and different variants on the transmission rate, I looked at data from Washington State specifically.

Figure 10 represents the new cases versus the total cases in Washington state from January 2020 to March 2022 plotted on a logarithmic scale. If the line increases linearly, it means that COVID is growing at an exponential rate. Otherwise, if the line dips down from its trajectory, it means that we have "flattened" the curve and slowed the exponential growth of COVID.

The green dots represent the time range from January to May 2021 when vaccines began being administered. The yellow dots represent August to October 2021 when the Delta variant became prominent. The red dots represent December 2021 to March 2022 when the Omicron variant became prominent.

As we can see, the introduction of vaccines drastically lowered the slope of the trajectory. The appearance of the Delta and the Omicron variant increased the slope again. However, the vaccines continued to be effective and decreased the slope of the trajectory sometime after each variant was introduced.

Interesting, the initial two dips in the trajectory of COVID transmission cannot be explained by vaccination. This may be the result of changing regulations, human behavior, or another factor.

4. How accurately can the number of fully vaccinated individuals be used for predicting new COVID-19 cases?

I initially used population, state area, and total COVID cases as features to predict new COVID-19 cases. Then I added the number of fully vaccinated individuals as a feature. I also used both a Decision Tree Regressor and a Neural Network to make predictions.

Unfortunately, none of the cases were able to produce a good prediction. I was unable to reach a high test score even after experimenting with different hyperparameters such as max depth of the Decision Tree and the layout of the Neural Network's hidden layer. All models seemed to be underfitting due to low scores on even the training data. In addition, the Decision Tree also seemed to be overfitting at time with an even lower score on the test data.

Including fully vaccinated individuals as a feature also did not increase the accuracy of the model. Out of 100 iterations of both the Decision Tree and the Neural Network, the Decision Tree performed better with vaccination data 45/100 times while the Neural Network performed better with vaccination data 40/100 times.

Therefore, I was unable to accurately predict the number of fully vaccinated individuals using population, state area, total COVID cases, and the number of fully vaccinated individuals as features. One reason could be that these features are not enough to capture the complexity of new COVID cases. More features, perhaps time, regulations, etc., might have been needed.

Impact and Limitations

None of the analyses performed in this document has rigorous statistical backings. I did not perform any statistical analysis, so all the connections and conclusions made should only be seen as correlations and not causations.

The data I was using can also be inaccurate. For example, states could have been reporting erroneous or intentionally inaccurate data that are then included in my study. Additionally, there are many factors with respect to COVID, such as immunity and state regulations, that I was unable to factor into this study. I also could not account for travel between states as another factor in contributing to new COVID cases.

However, there is definitive correlation between higher vaccination and lower COVID cases.

This study uses congregated data by states. The statistics of each state might not accurately represent the facts in each individual region or community. Even if a state has an overall low number of COVID cases, it does not reflect that all is well in every part of that state. For example, lower income communities and Native American communities might be disproportionately impacted by COVID and that might not be seen in analyses at a state level.

Although I was unable to improve the accuracy of predicting new cases by including vaccination data, it does not prove that vaccines do not contribute to lowering the transmission rate of COVID 19. Another model with more features and better hyperparameters still has room to improve accuracy.

Challenge Goals

Multiple Datasets: I will need to combine COVID-19 vaccination data and COVID-19 cases and deaths data to calculate the effectiveness of vaccines in each state or jurisdiction. I will also need to incorporate the population by state data, along with COVID-19 vaccination data, in order to make predictions about COVID-19 cases and deaths. To plot the data on a map, I will also need to incorporate US state boundaries data.

Machine Learning: I will be using machine learning to try to predict COVID-19 cases from COVID-19 vaccination and other factors. I will use both the DecisionTreeRegressor we learned in class and the more advanced Neural Network (Multi-Layer Perceptron) from the scikit-learn library to make predictions. I will experiment with different hyperparameters, such as the max depths of the Decision Tree and the layout of the Neural Network's hidden layers, to achieve higher accuracies. The main goal of using ML is to see whether ML models can improve the accuracy of predicting new COVID cases given the additional feature of the number of fully vaccinated individuals with all else staying constant. If the accuracy on the test set is improved after including the vaccination data, then there is a correlation between vaccination data and the number of new cases.

Work Plan Evaluation

The proposed workplan can be found in the appendix.

However, I made some major changes when carrying out my project. First, I decided to use the EdStem workspace instead of VS Code. Second, I choose to analyze thCole effects of vaccination on the COVID transmission rate instead of the actual efficacy of the vaccines.

Because I was using the EdStem workspace, setting up was a lot easier and only took a couple of minutes to do. My estimate for setting up the datasets was fairly accurate since it took me around 3 hours to do. Graphing the vaccine distribution and comparing each vaccine manufacturer (Figures 1-9) took a bit longer, around 5 hours, as it took time for me to finalize on the specific visualizations I wanted. Creating the plot of transmission rate for Washington State was not in my original plan and took around 3 hours as well. The main difficulty I had was plotting different colors on the same scatter plot. My estimates for the machine learning portion of my project were fairly accurate. It took me a very long time to finalize and extract the features and to settle on the hyperparameters I wanted to use. In my original plans, I also forgot to factor in testing, which took around 2 hours, and creating this report, which took around 4 hours.

Testing

I created a file, test.py, to run assert statements in order to test that all my functions for importing datasets were correct.

In order to test the visualizations, I ran internal visual tests. For example, I made sure that the graphs analogous to existing visualizations created by others matched those visualizations. I also added up the distribution of vaccines by each manufacturer and the result was equivalent to the overall distribution of vaccines.

To test the machine learning algorithms, I printed out the predictions and the labels or the true values of the test set to make sure they were at least close in value.

Collaboration

I determined the time range for when each COVID variant was the most prevalent in the United States here : https://usafacts.org/articles/covid-variants-delta-alpha-common/

I found answers to python syntax and specific library usage questions at StackOverflow: https://stackoverflow.com/

Appendix

Proposed Work Plan

Tasks

- Task 1: Download libraries and setup environment (2 hours)
- Task 2: Setup (download, import, clean, and merge) datasets (3 hours)
- Task 3: Graph distribution of vaccine administration (1 hour)
- Task 4: Compare vaccine types (2 hours)
- Task 5: Calculating and comparing the effectiveness of vaccines (5 hours)
- Task 6: Using Machine Learning to predict cases and deaths (8 hours)

Workflow

- Development: I will write out the main method pattern and function skeletons for each task first in order to layout the project. I will also comment each function before moving onto code development. Then, I will develop the code by going through the tasks one by one.
- Testing: I will test my code using the process we learned in class at least after each task and likely in multiple segments during each task.

Primary Development Environment: local development using VS Code and Anaconda